Silvia Hansen-Schirra, Stella Neumann, Erich Steiner

Cross-Linguistic Corpora for the Study of Translations

Text, Translation, Computational Processing

Edited by
Annely Rothkegel and John Laffling

Volume 11

Silvia Hansen-Schirra, Stella Neumann,
Erich Steiner

Cross-Linguistic Corpora for the Study of Translations

Insights from the Language Pair English-German

In collaboration with
Oliver Čulo, Sandra Hansen, Marlene Kast,
Yvonne Klein, Kerstin Kunz, Karin Maksymski
and Mihaela Vela

DE GRUYTER
MOUTON

ISBN 978-3-11-026029-8
e-ISBN 978-3-11-026032-8
ISSN 1861-4272

Library of Congress Cataloging-in-Publication Data
A CIP catalog record for this book has been applied for at the Library of Congress.

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.dnb.de.

Typesetting: RoyalStandard, Hong Kong
Printing and binding: Hubert & Co. GmbH & Co. KG, Göttingen
♾ Printed on acid-free paper
Printed in Germany

www.degruyter.com

Acknowledgements

This book reports on the outcomes of the project *Sprachliche Eigenschaften von Übersetzungen – eine korpusbasierte Untersuchung für das Sprachenpaar Englisch-Deutsch* ('Linguistic properties of translations – a corpus-based investigation for the language pair English-German') nicknamed CroCo (for Cross-linguistic Corpora) by a team fond of herding exotic fauna. It was funded by the German Research Foundation (DFG) as projects no. STE 840/5-1, STE840/5-2 and HA 5457/1-2.

We are greatly indebted to an anonymous reviewer for a detailed and very constructive report which helped us to clarify many points and to improve the structure of the current volume. The book would furthermore not have made it through the production process without Karin Maksymski's thorough and patient formatting work. Our heartfelt thanks go to her. We gratefully acknowledge our proofreaders' efforts: Paula Niemietz and Sarah Signer worked reliably and very fast.

Last but not least, we would like to thank the members of the CroCo team who collaborated with us on this volume: Oliver Čulo, Sandra Hansen, Marlene Kast, Yvonne Klein, Kerstin Kunz, Karin Maksymski, and Mihaela Vela, as well as several generations of diligent student assistants who helped compile and analyze the CroCo Corpus.

We are finally grateful to our editors at de Gruyter Mouton for supporting us in producing the volume. Needless to say all remaining errors and misconceptions are ours entirely.

The authors
Aachen, Germersheim, Saarbrücken
July 2012

Table of contents

Erich Steiner

1 Introduction

1 Topic

Our topic *Cross-linguistic Corpora for the Study of Translations: Insights from the language pair English-German* covers at least two major sub-domains:

On the one hand, we describe a corpus architecture, including annotation and querying techniques, and its implementation. The corpus architecture is developed for empirical studies of translations, and beyond those for the study of texts that are in some sense inter-lingually comparable, that is to say for texts of similar registers. The compiled corpus, *CroCo*, is a resource for research and is, with some copyright restrictions, accessible to other research projects.

On the other hand, we present empirical findings and discuss their implications for translation as a possible contact variety for the language pair English-German. Beyond our main focus on translation, though, our interest in the longer run is in language comparison and language contact more generally. The text property which is the focus of attention is *relative explicitness* of texts under comparison, and *explicitation* as a possible relationship between source texts and their translations in particular. *Explicitation* has often been assumed to be a specific property of translated texts, alongside possible other properties, such as *simplification*, *normalization*, *levelling out*, *sanitization*, *interference* and *shining through*. It is one of the motivations of the work reported on here to find out whether and to what extent the assumption of such properties can be supported through empirical work, and if so, whether these properties are interesting as influences on language contact phenomena.

Most of the research was undertaken as part of the DFG-Project *CroCo*, a corpus-based investigation into linguistic properties of translations for the language pair English-German.[1]

2 Motivation and goals

The long-term goal of our research is a contribution to the study of translation as a contact variety, and beyond this to language comparison and language contact more generally with the language pair English-German as our object

1 German Research Foundation (DFG) project no. STE 840/5-1, STE840/5-2 and HA 5457/1-2. For current information cf. http://fr46.uni-saarland.de/croco/.

languages. This goal implies, in our methodology, a thorough interest in possible specific properties of translations, and beyond this in an empirical translation theory.

The methodology developed is not restricted to the traditional exclusively system-based comparison, where real-text excerpts or constructed examples are used as mere illustrations of assumptions and claims, but instead implements an empirical research strategy involving structured data (the sub-corpora and their relationships to each other, annotated and aligned on various theoretically motivated levels of representation), the formation of hypotheses and their operationalizations, statistics on the data, critical examinations of their significance, and interpretation against the background of system-based comparisons and other independent sources of explanation for the phenomena observed. It is our belief that over the past couple of years sufficient progress has been made in corpus technologies and in extracting information on the data to render such an endeavor promising.

3 Theoretical foundations and state of the art

Theoretical foundations of the developments outlined here are to be found
- in the more textually-oriented and linguistically-based strands of translation studies (3.1),
- in models of linguistic variation and register (3.2),
- in the area of corpus design and implementation, and corpus technology more generally (3.3),
- in studies of language comparison and contact, with a focus on language-specific ways of encoding meaning (3.4).

This introduction aims at an outline of the theoretical foundations on the most general level only, because individual chapters will review their own locally relevant state of the art. However, there are some theoretical foundations which form a sort of macro-background for our overall enterprise, and it is this general background which will be sketched here.

3.1 Translation studies

There is a tradition of assumptions *in the more textually-oriented and linguistically-based strands of translation studies* about specific properties of translated texts. According to such assumptions, translations are characterized by specific textual

properties; they constitute a "text-type", or "register", of their own (cf. Frawley 1984; Blum-Kulka 1986; Sager 1994; Toury 1995; Baker 1993, 1996; House 1977, 1997, 2002, 2008; Steiner 2001a, 2001b; Teich 2001, 2003; Hansen 2003; Neumann 2003; cf. Fawcett 1997: 100 and Laviosa-Braithwaite 1998 for overviews). These assumptions, and some hypotheses deriving from and specifying them, have been subjected to some initial empirical testing, but nothing approaching an accepted answer to the question embodied in it has been found to date. Furthermore, where some properties of translated texts have been tentatively identified so far, no consensus is in sight as to whether such properties might be mainly due to the specifics of the translation *process*, and in that sense universal to translations, or whether they must rather be explained by recourse to contrasts between the linguistic systems involved and/or by contrasts between the text types, or registers, of the source and target texts and specific translation strategies deriving from those.

Translation studies and linguistics have produced a body of work on language pair-specific and sometimes direction-specific translation problems and translation procedures which provides valuable initial insights on implications of language contrast for translation (Vinay and Darbelnet 1958/1995 in their comparative stylistics of English-French; didactically motivated explorations for the language pair English-German [Friederich 1977; Purser and Paul 1999; Königs 2000], more linguistically founded work by Doherty throughout the 1990s culminating in Doherty 2002 and 2006, and differently House 1977, 1997, both for English-German mainly, or Fabricius-Hansen 1996, 1999 for the language triangle English-German-Norwegian). These studies contribute significantly to our understanding of language-pair specific processes and relationships in translation, without, however, foregrounding the question of whether there are "universal" properties of translated texts. Neither are they methodologically empirical in the stricter sense. By "in a stricter sense" we mean, initially, based on a somewhat larger quantity of data, sampled with some technique aiming at representativeness, and using categories of data which allow a transparent relationship to research questions formulated, and also repeatability of the analysis by different researchers at different places and times.

More recent years have seen the emergence of empirical investigations into universal properties of translations (Baker 1996; Laviosa-Braithwaite 1998; Olohan and Baker 2000; Kenny 1998, 2001; Olohan 2004; cf. House 2008 for a critical overview), where the assumed properties were of the type *simplification*, *normalization*, *levelling out*, *sanitization*, *disambiguation*, *conventionalization*, *standardization*, *avoidance of repetition* and in particular *explicitation* (cf. various contributions in Mauranen and Kujamäki 2004; Saldanha 2008; Englund Dimitrova 2005; and for an earlier summary Klaudy 1998). The property of *explicitness*

and the process of *explicitation* will be defined and operationalized in some detail in chapter 4. About the other properties, we would like to say a bit more at this point. *Simplification* usually refers to increasing "readability" of a text, for example by simplifying a type of linguistic structure, e.g. in terms of number of constituent elements of some linguistic unit. Other measures include increased and more explicit punctuation, decreased lexical density or decreased type-token-ratios. *Normalization* refers to a process within which a (translated) text approximates or even exaggerates some norm of the target register it is translated into, always in terms of some selected textual/linguistic feature. Normalization also often means the avoidance of some syndrome of marked features or structures in target texts. *Levelling out* is always predicated of sets of texts, for example when we hypothesize that a set of translated texts, when compared to a set of non-translated texts of a given language and a given register will be composed of texts which are more similar to each other in terms of some (set of) linguistic features, in other words, the range of variation among translations are assumed to be smaller than for otherwise similar original texts. *Sanitization* as a property is assumed to be given when translations avoid affectionally strong language, in particular stigmatized language, relative to original texts. *Shining through* in the sense of Teich (2003: 209–218) means an interference in a translation from its source language, but often in terms of proportionalities and frequencies, rather than simply in terms of individual structures or lexical items as in cases of simple "interference".

We shall meet these, and other, assumed properties of translations as phenomena to be tested throughout our study (especially chapters 5ff.), even though usually our emphasis is on investigating explicitness and explicitation. As far as the assumption of universal properties of translations is concerned, though, our general stance is probably close to the cautious and skeptical attitude adopted in House (2008: 10–12): Much of what is all too loosely postulated as a "translation universal" may well turn out to be either a general property of language (use), or it may be specific for some given combination of languages, it may be specific to one direction between two languages, it may be strongly dependent on register or genre, it may be sensitive to language-change phenomena. In any case, whatever there may be of translation universals, it could be restricted to a highly general level only: one such highly general "universal" may be the fact that each translation necessarily represents an attempt at optimizing conflicting constraints posed by the ideational, interpersonal and textual functional dimensions of encoding – which would be a universal so general that its predictive power would be very limited, unless it were reformulated as much more specific instantiations of that general assumption – something which we believe to be possible in principle. However, and maybe slightly more "universalist" than the

stance adopted in House (2008), if it could be shown that an assumption about de- and re-metaphorization in translation-oriented psycholinguistic processing of the type made in Steiner (2001a: 170ff., 2001b: 15ff.), Hansen (2003: 118–125), and summarized again here in chapters 7 and 14, is valid, then this could be the source of a property shared by all translated texts relative to non-translated ones, even though the kind and extent of explicitation would be strongly sensitive to language-pair specific and direction-specific factors. We have begun investigations of such de-/re-metaphorization processes in process-oriented experiments (Alves et al. 2010) in which we focus on interactions between variable translation-units (in a processing sense) and degrees of metaphorization, where "metaphorization" is always to be understood as "grammatical metaphor" in the sense of Halliday's "Functional Grammar" (Halliday 1985: 319; Halliday and Matthiessen 2004: 586; see also chapter 7).

However, independent of whether or not any of the assumed properties of translated texts are general across more than two languages, genres, registers, we see their particular research potential in their relationship to feature- and property-based approaches, to contrastive linguistics, language contact studies and issues to do with processing. Languages and texts can usefully be contrasted in terms of properties; they can be assumed to influence each other in terms of such properties. Dynamic processes such as language change and language processing can be modeled on properties – and we would hope to be able to interface with empirical research traditions currently being developed in these areas, some of which we shall address below, and again towards the end of this book.

The line of argumentation positing properties of translated texts, even though we discuss some aspects of it critically here, represents progress towards an empirical research methodology, as well as an increased focus on properties of translated texts due to the translation process. While it thus has paved some of the way for our own goals, some of it suffers, in our view, from impoverished linguistic modeling: its essentially corpus-driven, rather than theory- or model-driven, methodology and the linguistically low level at which phenomena are operationalized make it very difficult to address higher and more theoretically meaningful linguistic levels, lexico-grammar, semantics and text/discourse in particular. It is therefore also no coincidence that within this line of research, the valuable insights of language typology and typologically-based linguistic comparison are not exploited in explanations of the phenomena observed.

So far, then, we are claiming that on the one hand, the linguistically more informed studies of translations mentioned above would gain from a more empirical methodology, and from taking the process of translating as a mode of text production more seriously as a source of explanation. On the other hand,

existing and methodologically more empirical studies of translations would need much more of an influence of linguistic models of variation and register, and of studies of language comparison and contact, with a focus on language-specific ways of encoding meaning, in order to be able to make a contribution not only to our awareness of isolated and theoretically sometimes arbitrary features which characterize translations, but rather to our understanding of translations as texts, and to translations as a possible contact variety between languages. Both research strands could gain substantially from devoting more explicit attention to the areas of corpus design and implementation. In these areas we hope to be able to make a contribution, and we would like to start with computational design and implementation of corpora, before turning to the linguistic basis for the modeling to be suggested here.

3.2 Models of linguistic variation and register

In terms of general awareness of tools and architecture in corpus technologies, we are, like many other projects, indebted to *models of linguistic variation and register* (cf. Biber 1988, 1995; Biber, Conrad, and Reppen 1998) and to work on languages in contrast (cf. the SPRIK project in Oslo, for example Johansson and Oksefjell 1998). As part of this legacy, we have attempted to integrate statistics for the evaluation of the significance of results where appropriate (cf. Biber, Conrad, and Reppen 1998; Butler 1985; Oakes 1998).

As for models of linguistic variation and register, we obviously need an understanding and some modeling of how, and along which dimensions, texts can be classified as similar or different. A "lean" variant of such a model is the notion of "register" as used in Biber's work, or in Biber et al. (1999). A richer and theoretically more committed variant is the notion of "register" in its original theoretical context in Systemic Functional Linguistics (cf. Halliday, McIntosh, and Strevens 1964: 87–88; Halliday and Hasan 1989; Matthiessen 1993). Translation studies have a substantial history of using this notion (cf. House 1977, 1997: 196; Hatim and Mason 1990; Hansen 2003: 23; Neumann 2003: 16; Steiner 2004b: 11), and we have used it in various degrees of theoretical commitment (for an advanced example cf. Neumann 2008). In a "lean" version, register theory can be seen as not much more than some form of text typology, and quite a few of our studies use it just in this "lean" version. In a more theoretically-committed version, the dimensions of variation of this typology systematically link up with the linguistic system and its multi-functional grammar on the one hand, and with the context of culture on the other. The modeling translation within this overall architecture can be seen in Matthiessen (2001), Teich (2001) and Steiner (2001a).

3.3 Corpus design and implementation

In the area of corpus design and implementation, we have imported and further elaborated techniques from multi-layer corpus architectures, annotation, treebank technologies and information extraction on data in such corpora. A fundamental characteristic of our methodology is that we are not working on raw corpora, but on multi-layer annotated corpora (with and without alignment), bridging the gap between the formulation of hypotheses on higher levels of linguistic structure and their operationalizations in instantiated texts (cf. Hansen 2003; Teich 2003; Neumann 2008).

On a more technical note, existing corpus tools have been used – ranging from automatic to semi-automatic to computer-assisted manual annotation and alignment (cf. Lüdeling and Kytö 2008, 2009 for an overview). These include some tools that are language-independent, but the trade-off for the high degree of flexibility is a low degree of automation. Other tools enabling automatic or interactive annotation require language-specific training, which raises the question of comparability across multilingual annotations (cf. Neumann and Hansen-Schirra 2003).

The multi-layer annotation and alignment of the CroCo Corpus allows us to view the annotation in aligned segments and to pose queries combining different layers. The resource thus permits the analysis of a wealth of linguistic information on each level helping us to understand the interplay of the different levels and the relationship of lower-level features to more abstract concepts. For this purpose, two technical requirements must be met: the exploration of the integrated data (i.e., simultaneous viewing of the different levels and searches across levels) and integrated processing, e.g. for the discovery of correlations across layers. These requirements are met by using stand-off annotation at each layer on the one hand (cf. McKelvie et al. 2001) and alignment of base data across the layers on the other (Bird and Liberman 2001). Developed for multi-layer annotation in XML, the XML Corpus Encoding Standard (XCES) guarantees exchangeability and consistency since predefined XCES Schemas, DTDs and XSLT scripts can be used (Ide, Bonhomme, and Romary 2000). For efficient querying, the annotation and alignment information can be stored in a relational database (cf. Cassidy and Harrington 2001), which allows the integration of hierarchical annotation layers. Chapters 6–11 will show that empty alignment links, crossing alignment lines as well as the combination and exclusion of annotation tags are important for the linguistic exploitation of the CroCo Corpus. The results of such combined queries can then be interpreted in terms of linguistic properties of translated text.

3.4 Studies of language comparison and language contact

Let us now turn to *studies of language comparison and language contact*, with a focus on language-specific ways of encoding meaning:

Language contact is the situation in which languages, or rather, instantiations of language systems through their speakers, influence each other synchronically in shared socio-semiotic contexts (classical accounts include Weinreich 1953; Thomason and Kaufman 1988; Oesterreicher 2001; a more recent account is given in Siemund and Kintana 2008). This is complementary to the historical axis, along which genetically related languages are in contact through time. Language contact applies to varieties within languages, as it does to different standard languages. Major topics of research are (cf. Thomason and Kaufman 1988: 65–100):

- the interplay between synchronic contact and genetic inheritance
- linguistic vs. socio-cultural constraints on interference
- analytic frameworks for contact-induced language change (linguistic levels of change; borrowing vs. interference through shift; predictive power of the frameworks, external vs. internal explanations)
- language maintenance
- normal vs. exceptional transmission (creolization, pidgins)

In an attempt to generalize on the strength and on linguistic levels of language contact, a *borrowing scale* is postulated, ranging from lexical borrowing only through slight structural borrowing, moderate structural borrowing and finally to heavy structural borrowing. Most studies to date have focussed on lexical items and/or grammatical structures, rather than on features or properties of the linguistic systems and instances (discourses, texts) involved, although both perspectives have often been acknowledged as relevant (cf. also Heine 2008: 37 in his Figure 1 on contact-induced linguistic transfer).

Multilingualism is usually predicated either of individuals, or of linguistic communities as socio-cultural formations, or else of discourses/texts (for a representative and comprehensive survey cf. Auer and Wei 2007). In the first sense, studies of multilingualism are often carried out as studies of language development/acquisition of several languages in one speaker (bilingualism, trilingualism, etc.). In the second sense, they are targeted at linguistic communities and are methodologically situated in sociolinguistics. A terminological distinction which reflects this division is that between *bilingualism* as referring to the individual, and *diglossia* as referring to communities. In the third sense, there are a few strands of research into *multilingual text production* (cf. Matthiessen 2001; Steiner and Yallop 2001; Teich 2001; Steiner 2004a, 2004b, 2005a, 2005b, 2005c), cross-cultural pragmatics (House 1997, 2002) and information structure

across languages (Hasselgård et al. 2002; Fabricius-Hansen and Ramm 2008), in which *multilingualism* is treated as a property of discourses which are assumed to have interesting and typical properties compared to *monolingual* discourses (cf. several contributions in Franceschini 2005, in particular von Stutterheim and Carroll 2005). If we say that *discourses* are multilingual, then we imply that they show special *discourse properties* of *directness vs. indirectness*, *orientation towards self vs. other*, *orientation towards content vs. interaction*, *explicitness vs. implicitness*, *routine-orientedness vs. ad-hoc formulation* (as e.g. in House 1997: 84), ultimately to be realized in lexico-grammatical phenomena such as interference, borrowing, code-/language switching, special metafunctional orientations in terms of ideational, interpersonal, or textual biases, directness, density, explicitness, and others. These discourses thus instantiate specific contact varieties, or registers. In our own research, we regard translations as an important venue of influence in language contact (cf. Frawley 1984; Baker 1996 for translations as a special text type or even code). But this venue of influence is additional to, and different from, more traditional venues of contact through borrowing or interference. It is less obvious, the resulting varieties are superficially close to native ones, and it applies intra-lingually, across registers, as much as it does inter-lingually.

Investigations of *multilingualism* are meaningful on all of the levels mentioned above, provided the empirical claims that are being made by the ascription of the property to individuals, communities, or discourses are clear. Furthermore, in the case of discourses, it must be clear whether empirical claims are made about properties on the level of text/discourse, or else on the level of lexico-grammar – or about both of them. A multi-functional and feature-based perspective will usually encompass the discourse-oriented perspective, at least as an important component, and certainly as a prominent object of study. *Multilingualism* of discourses can be assumed to be a property which is both a result of, and an environment for, language contact and change.

In a first attempt to characterize our own research efforts relative to the substantial tradition of research briefly characterized so far, it will be obvious that they rely for their modeling to some extent on Systemic Functional Linguistics (cf. Halliday and Hasan 1976; Halliday 1978; Halliday and Martin 1993; Halliday and Matthiessen 2004). We have additionally drawn on comparative and typological perspectives with some functional leanings (cf. Hopper and Thompson 1982; Hawkins 1986; Thomason and Kaufman 1988; Biber 1995; Simon-Vandenbergen and Steiner 2005; Traugott and Dasher 2005) and on insights from certain strands in translation studies, contrastive linguistics and cross-cultural pragmatics (Doherty 1996, 2002, 2006; Fabricius-Hansen 1996; House 1977, 1997, 2002). In

terms of methodology, a perspective of the kind advocated here will give due consideration

- to systems alongside structures,
- to the instance alongside the system,
- to more abstract types of contrast, for example in terms of *explicitness*, than have often been in the center of theorizing, and
- to the metafunctional modularization of language.

Corpus-based work in our own group on original and translated texts in English and German (cf. Hansen 2003; Neumann 2003, 2008; Steiner 2004b; Teich 2003) shows how an instance-based orientation of work on multilingual discourses can yield new insights and methodologies in addition to the more traditional system-based investigations. It is also within this instance-based perspective that properties of discourses come into view which are below the threshold of consciousness of language users, and outside the realm of borrowing of lexical or structural patterns across languages. A typical case are "good" translations, which often show no lexical or structural trace of language contact, but which may have a characteristically different "feel" to them, which is the result of different frequencies and proportionalities of native patterns, rather than the result of borrowing or interference on the lexical or structural levels. What these pieces of research do, methodologically, is to combine a Biber-type corpus orientation (cf. Biber 1995 and elsewhere) with multi-layer corpus architectures and annotations, elaborated querying techniques and modeling of multilingualism against the background of more structured linguistic theories, especially of functional orientations.

We have furthermore attempted to derive from lexico-grammatical patterns some more abstract (and at the same time, more empirical) properties than have often been in the center of theorizing (Steiner 2004b, 2005a, 2005b, 2005c, 2008b; Hansen-Schirra, Neumann, and Steiner 2007). One of these, *explicitness*, is in focus here. Languages, through their instantiations in texts and discourses, influence each other in contact situations if there is some relevant sense of a contrast. Traditionally, these contrasts have often been sought in the non-existence of lexical items and their immediate grammatical environment in a receptor language. Beyond the lexical level, borrowing scales such as the one postulated by Thomason and Kaufman (1988: 65–100), are an attempt at systematizing processes of borrowing into receptor languages, or interference from source languages, in terms of grammatical structure. There is an underlying assumption of gaps in the receptor language, or otherwise a strong influence of the source language through *shifting* speakers. Again, the expectation is one of some relevant contrast inviting the borrowing or interference, which is an expec-

tation shared in our work. However, we posit additional levels of observation and modeling: in the first place, the relevant contrasts may be in terms of higher-level text- and discourse structures, and only through these in lexis and grammar. In the second place, the contrasts may manifest themselves initially and for a substantial period in terms of changing frequencies of existing lexico-grammatical configurations, rather than in the borrowing or interference of "foreign" contact-induced lexico-grammatical structures. Pressure towards language change builds up, as it were, through changed frequencies of existing constructions long before it manifests itself in new structures on any one of the linguistic levels. This does not mean that any of the more traditional studies of language contact and change are unimportant or obsolete, but rather that perceived differences in ways of structuring discourses often have to do with changes in relative frequencies. These are then perceived as making a text more or less *explicit*, *direct*, or *dense* than some received norm in some language or variety within a language. These differences are often hardly above the threshold of perception, and thus constitute much more of a cline of perceived properties of texts/discourse, than coarse binary distinctions such as *native vs. non-native* command of a language or variety would suggest (cf. Franceschini 2005 and several contributions therein). And in this sense, translations can be expected to constitute a prime example of contact varieties.

Finally, several linguistic frameworks have postulated a modularization of linguistic structures along different dimensions, usually adopting some diversification into, roughly, referential/ideational/propositional vs. interactional/interpersonal vs. textual/organizational meanings. The latter dimension is an *enabling* function yielding structure in terms of Theme vs. Rheme, Topic vs. Comment and Given vs. New information. A model giving prime architectural place to these distinctions is Systemic Functional Linguistics (SFL). Within such a model, avenues of language contact will be modularized by metafunction, and will be conceptualized to operate on properties (features), rather than on structures primarily. The more structure-oriented tradition in grammaticalization studies has focused on explorations of morpho-syntactic change, building on Lehmann's (1995) classic study on processes and parameters of grammaticalization. This type of grammaticalization research mainly focuses on the change of free syntactic units into highly constrained morphemes with a grammatical function. A significant step in the more system-based and multi-functional direction has, outside of SFL, been taken in some more recent work by Traugott and Dasher (2005: 19–24, 81–88), who focus on semantic-pragmatic change in grammaticalization. They hypothesize semantic change to proceed along the following cline: *propositional towards textual towards expressive*. A cline such as the one postulated here is, of course, strongly reminiscent of the metafunctional

modularization in SFL of dimensions of grammatical structure into ideational, textual and interpersonal, an architecture which has been exploited in recent typological work on a range of languages (cf. Caffarel, Martin, and Matthiessen 2004; Steiner and Teich 2004; and Matthiessen 2004 in particular).

Finally, the issue should be raised of how the notions of *explicitness*, and ultimately also *density* (cf. Bickel 2003; Noonan ms.) and *directness*, alongside the frequently employed notions of *directionality of change* and of *frequency of usage* (Bybee and Hopper 2001: 1–2) may have a bearing on models of language contact and language change operating on properties of encoding in a multi-functional and feature-based view on language.

Our point of departure will be a working definition of the notions of *explicitness* and *explicitation* as discussed and defined in chapter 4 below. *Directionality of change* and *frequency of usage* may then have implications for a modeling of language contact (and change) in terms of explicitness and related properties.

Directionality of change is a notion which can be predicated on different types of structure. In earlier versions of that notion, we encounter hypothetic developments from morphologically synthetic to analytic language types, between types of basic word order, or between types of marking relations such as head-marking, dependent marking, mixed-marking, etc. Some influential work has postulated cycles of development, driven by the dialectical needs of language users towards increased expressiveness on the one hand, and maximal economy on the other (e.g. Hagège 1993: 147–148). In more recent times, directionality has sometimes been linked to multi-functional, or multi-dimensional models of language, as for example in the work of Traugott and Dasher, who refer to Halliday's multi-functional hypothesis (Traugott and Dasher 2005: 94–95). Halliday and colleagues have, indeed, in several places raised the issue of language change (e.g. Halliday and Matthiessen 1999: 227, 507; Matthiessen 2004: 655), frequently in connection with ideas from general systems theory, without so far having spelled out all the implications.

Traugott and Dasher trace a line of theorizing which assumes interactions between *subjectivity*, *intersubjectivity* and *objectivity* in language (use). In earlier versions, Traugott and Dasher (2005: 94, but originally in Traugott 1982) had postulated a unidirectional development of semantic change along the lines of *propositional* > *(textual)* > *expressive*, which later on they differentiated into sub-types (Traugott and Dasher 2005: 281). Very interestingly, *subjectivity*, *intersubjectivity* and *objectivity* seem to be properties of grammatical constructions, much in the same way as we conceive of *explicitness*, *density* and *directness*. Where we see our role relative to this interesting line of research is in a comparison of the kinds of abstraction we are making, in their relationship to the multi-functional hypothesis, and, importantly, in our attempts at developing

empirical research methodologies based on electronic corpora. We also aim to trace the contribution of situations of multilingualism and translation to language contact and change. Finally, we would like to investigate particular registers as sites of contact and change (cf. Traugott and Dasher 2005: 283–284, also their remarks on historical pragmatics and historical discourse analysis 99; importantly House e.g. 2002). What needs to be clarified is the precise locus of the phenomena we are talking about: grammar, semantics, discourse, or the mapping between them (Traugott and Dasher 2005: 282–283). One attempt at this clarification is made in our remarks about explicitness and explicitation in chapter 4 of this book (for more detail cf. Steiner 2005c; Hansen-Schirra, Neumann, and Steiner 2007; but also Doherty 2006: 49–50).

Staying with *explicitness* for a moment, we would speculate that it is a property of constructions and configurations both on the textual and on the lexicogrammatical levels. We would furthermore like to suggest that translations and other forms of multilingual discourses show degrees of explicitness differing from the explicitness of encoding in registerially related non-translational or otherwise monolingual discourses. And we would also assume that it is partly these differences through which pressures towards expressiveness or economy exert their force, thus becoming driving forces of change. Such change, though, would not simply presuppose the existence of relevant differences in explicitness and other properties of that type, but also certain critical levels of frequency before they become effective.

Frequency of use can be found in several of the studies in Halliday (2005: 93), where it is argued that linguistic sub-systems can be more or less stable as a consequence of proportional frequencies between relevant types of construction, for example the relationship between positive and negative clauses in the environment of primary tense in a big corpus. These frequencies, and the resulting *markedness*, may be among the driving forces of change (cf. also Johanson 2008: 74–75 and his notion of "frequential copying"). In Bybee and Hopper (2001: 2–3), frequency is accorded a key role in the *emergence* of structure in discourse (cf. in particular MacWhinney 2001: 449–450, 464–465). It affects the strength of a pattern, it works differently on types, tokens and collocations, has effects on pattern productivity, may preserve old structures, it may positively work towards fusions, contraction and affixation, it may increase accessibility to (sound) change, or it may increase accessibility to semantic bleaching and other functional changes. Particularly within an approach operating with *properties* of constructions alongside the constructions themselves, frequency of a construction may itself affect properties (such as explicitness). And finally, the frequency of more or less explicit discourses may be a driving force in change. All of these processes can be assumed to be influenced by the degree of multi-

lingualism of discourses, and this is what we would like to explore in more detail.

It is hoped that new investigations of the phenomena addressed in the research strands just mentioned, but with a stronger basis in empirical and corpus-based techniques, will enable us to critically examine available views and add new perspectives based on a systematic investigation of more data, and also more structured data than was possible before. We would very much like to contribute towards empirical and corpus-based techniques with our work reported on here.

4 Methodological principles of the studies

Specific questions of methodology will be discussed in the relevant chapters. On a global level, a corpus architecture will be described which specifies *types of contrast* investigated in the corpus, the levels and specificity of the *linguistic phenomena covered*, and the *results* and *kinds of explanation* which can be evoked against the architecture of our corpus and against independent sources of explanation (language type, register, translation as a process of text production).

Without going into much detail here, we would state our overall commitment to narrowing the gap between high-level hypotheses and data: texts are often assumed to be of different degrees of *explicitness*, *density*, *directness*, *simplicity*, *addressee orientation*, *content orientation*, *objectivity*, *subjectivity*, etc. These assumed properties in terms of which texts are often compared cannot, however, be read-off from the data directly. They need to be operationalized in terms of linguistic properties of constructions on various levels, such as lexico-grammar, cohesion, "epiphenomenal" properties of entire texts, such as lexical density, type-token relationships, part-of-speech profiles, etc. A first type of investigation, which we will call "descriptive", then attempts to locate significant contrasts between texts and sub-corpora in terms of such linguistic properties and their interactions (see for example chapter 5 of this book). General assumptions are furthermore operationalized and specified into hypotheses from which we derive queries, and the results of these can then be used in attempts to falsify the hypotheses. As usual, this is a multi-level process of data and interpretation, but a process in which we would like to motivate our interpretations as closely as possible by the relevant level of data. Whether or not, for example, a given piece of text is more or less explicit than another needs to be interpreted in terms of morphology, types of words, types of phrases, types of clauses, and types of cohesive patterns – and always in terms of proportionalities and relationships between them. This process is still one of interpretation, the data do

not "speak for themselves", but the interpretation is heavily constrained on each of the levels involved by our operationalization of what *explicitness* and *explicitation* is. But even before this stage is reached, the annotation of the data, say in terms of parts-of-speech, is, of course, a process of interpretation already, and again one which needs to be as tightly operationalized as possible.

We also need to be able to contrast and compare different types of corpora: reference corpora with any register-specific corpora, register-specific corpora with each other, both intra-lingually and inter-lingually, translations with originals within one language, and across languages as source vs. target texts, and, importantly, we need to be able to investigate translation units in aligned corpora. Secondly, we need to be able to make comparisons on very different levels: on the lexico-grammatical level, and within that on all the ranks from morphemes up to clause complexes, as well as on the text level when we are investigating cohesion in all its different forms. And finally, we need to be able to check our results against possible sources of explanation (language type, register, translation as a process of text production, possibly others), which means we must have ways of grouping independent and dependent variables with the ultimate aim of tracing causal relationships between them – if possible. In other words, our overall corpus consists of a number of sub-corpora which can be grouped into various constellations, all the corpora are annotated on a series of linguistically motivated levels, and in general, we want to be able to move from description – which is interesting in itself – to explanation.

While the assignment of a text to German or to English as languages is relatively clear for data collection in our case, the further distinction as to the register into which a given text sample belongs is much more difficult. This methodological question is one we share with all projects using register-specific data and will be addressed in more detail in chapters 2 and 3. However, our approach has to face the additional question of what counts as a translation and what counts as an original. We have adopted here the relatively "open" strategy of admitting any text as a translation which was produced and published as one. This of course means that our translation sub-corpora include samples which are on the borderline between translation and multilingual text production.[2] Maybe more significantly even, they contain "translations" which are clearly non-optimal and/or even contain errors and mistakes. When we are therefore making statements about such texts, these may be partly due to such "impure" phenomena. Some authors, for example Doherty (e.g. 2006: 1–2) have

2 Translations being pairs of source-texts and target-texts, whereas multilingual text production refers to cases where texts are produced in parallel in different languages from some knowledge source other than a linguistic source text.

argued that, if we want to investigate translation as a mode of text production, we need to work on evaluated data, on other words on "good" translations. And ultimately, we share the interest in texts which are motivated as "translation" as part of some model, rather than on arbitrary texts called "translations" just in terms of some ill-informed socio-cultural labeling. However, as a *detailed* model of translation can only be the result of an empirical endeavor – as opposed to an *initial and general* model, which has to actually guide empirical work from the start – we believe we cannot afford to rely on intuitions about "good translations" too early on. Furthermore, language contact happens through texts which are actually out there and are being processed as such – even if they are imperfect. We therefore adopt a relatively liberal strategy in admitting texts into the corpus, but will of course ultimately want to say whether and above all why and how some text is or is not a "good" translation.

5 Road map

The book is organized as follows: Part I will introduce our corpus resource in accordance with the methodological principles described above. These include the design criteria of the corpus, the automatic and manual annotation, alignment on the levels of word, chunk[3], clause and sentence as well as the technical specification needed for this corpus design. We will describe how the quality of the linguistic enrichment is ensured and how the resource can be queried. A final topic in this first part is the combination of qualitative and quantitative investigations in the study. The corpus resource thus described is available to other researchers and other types of research questions than our own, even if for legal reasons, the corpus itself can only be accessed locally at this stage.

Part II addresses some findings about *explicitness* and *explicitation* as relevant properties of contrastive text corpora, emerging from the exploitation of the resource. First, we develop hypotheses about explicitness, which are then operationalized in terms of indicators of explicitness and explicitation. This is followed by a characterization of the resource and the types of contrast which can be investigated based on *shallow* statistics, by which we mean lexical density, type-token-relationships and part-of-speech (PoS) proportionalities within and between the sub-corpora. The overall aim in this part is to arrive at profiles for

3 'Chunk' is the cover term used in the CroCo project for intermediate grammatical units. It covers both the formal interpretation in terms of groups/phrases as well as the functional use in terms of subject, object, predicate etc.

the various types of contrast in the corpora (between languages, between registers and between originals and translation). These profiles will also receive some initial interpretations. The next step is a heuristic examination of translation properties with a view to guiding further hypothesis formation. Against the background developed up to that point, a small number of case studies in the corpus will be reported on in Part III. These will cover shifts in grammatical functions, in particular the Subject function, but also the class of Adverbials and their position in linear order, then shifts in (co)reference, shifts in cohesive devices in translated texts, an investigation of information distribution in English-German noun phrases and their shifts in translated text, and finally register-induced properties of translations. These case studies are intended to illustrate the types of findings which we can expect from the methodology developed earlier on. Chapter 14 later on in Part IV will move over into the area of explanations. Possible explanations are derived from systemic contrasts between English and German (cf. Rohdenburg 1990; Hawkins 1986; König and Gast 2007/09; Steiner and Teich 2004), from register, and from the nature of the translation process.

Part IV discusses computational perspectives of the CroCo Corpus. Here, the potential of the corpus as a parallel treebank as well as its limitations are revisited. We will additionally give an outlook on computational applications of the resource beyond our immediate goals. This includes, for instance, the development of an API and a bilingual gold standard as well as the usability of the corpus for machine translation and other tasks in computational linguistics.

The book is rounded off by generalizations, conclusions and outlook (Part V) addressing the research questions mentioned above, which in translation studies have so far mainly been discussed in an intuitive rather than empirical way.

I Texts – The CroCo resource

Stella Neumann & Silvia Hansen-Schirra

2 Corpus methodology and design

1 Introduction

The present chapter discusses aspects of corpus-based linguistic research as one type of empirical research. After some theoretical considerations of empirical linguistic research, we will introduce the specific design chosen for the CroCo Corpus. This chapter and the following chapter 3, which covers the more technical aspects of the CroCo resource, give an overview of the methodology of the CroCo project.

2 Theoretical considerations

Introductions to corpus linguistics typically start by discussing the difference between empiricist and rationalist approaches to the study of language.[1] This has been widely discussed (cf. for example the contributions in Svartvik 1992) and is not the major concern of the present chapter.

The corpus approach investigates naturally occurring language and is thus intrinsically empirical. "Empirical method" refers to the research method which investigates actual data. In this sense, "*empirical* indicates that the information, knowledge and understanding are gathered through experience and direct data collection" (Black 1999: 3). One of the main characteristics of the empirical method is that it allows *systematic* observations with the goal of producing replicable studies (Black 1999: 4). Halliday (2005: 173) refers to corpus linguistics as an empirical approach to the description of language where the accumulation of new data and their interpretation leads to new theories. He states: "after all, that's what it did in physics, where more data and better measuring transformed the whole conception of knowledge and understanding. How much

1 See for instance chapter 1 of McEnery and Wilson (2001), chapter 1 of Meyer (2002) and chapter 2 of Lemnitzer and Zinsmeister (2006). Featherston (2008) exemplifies a systematic approach to an intuition-based investigation of language that overcomes the typical critique by empirical linguists of its introspective and consequently non-systematic character. It does not, however, overcome the non-naturalness of isolated and possibly artificial sentences (see Chafe 1992: 86).

the more might we expect this to be the case in linguistics, since knowing and understanding are themselves processes of meaning."

There are, however, some philosophical issues associated with empirical research which should be kept in mind when evaluating the explanatory power of empirical findings (cf. Neumann 2008). Referring to Thomas Kuhn's work on "scientific revolutions", Okasha (2002: 88–89) explains the "theory-ladenness of data": as a matter of perception, different people look at data from different theoretical perspectives and thus perceive the data to be different. While this statement appears to be confirmed by day-to-day experience in scientific discourse over linguistic findings, Okasha qualifies it by pointing out that this does not rule out objectivity altogether since scientists from different paradigms may accept certain statements that are "sufficiently free of theoretical contamination" (Okasha 2002: 89). Despite their different opinions on whether there is such a thing as objective findings, most philosophers of science will accept the existence of an objective truth. Efforts aimed at ensuring objectivity are concerned with whether a study produces the same results irrespective of the person by whom the analysis is carried out. The more a given study relies on human interpretation, the more important this concept becomes. Typically, in qualitative studies the concept of objectivity is replaced by intersubjective veryfiability ensured by transparent documentation of the research process, the use of codified procedures (in linguistic analysis this is achieved, for instance, by adhering to clear annotation guidelines) and transparent data interpretation.

Other general concepts aimed at ensuring the quality of empirical research are *reliability* and *validity*. Reliability is concerned with the exactitude of the measuring instrument or method. If the instrument produces accurate results, repetitions of the study under the same conditions should yield the same results. Apart from systematic errors due to the limits of accuracy of automatic tools (which may indeed make a tool useless for linguistic rather than computational linguistic research), reliability can be one advantage of using tools, assuming that they do not change their interpretation of a given element (which could, however, happen with purely statistical tools) and that – unlike the human analyst – they do not get tired. Validity refers to whether the choice of method is appropriate to the phenomenon under investigation and whether the chosen indicators actually measure the concept under investigation (and not a confounding factor). This is of particular importance in quantitative studies relying on hypothetical relations between the abstract concepts of interest and the linguistic indicators used to obtain information on the concepts.

Apart from these quality criteria used in the social sciences, another evaluation method is of relevance to the automatic processing of corpora in general. A measure used to evaluate the success of natural language processing (NLP)

systems, especially in information retrieval, is *precision and recall* (Manning and Schütze 1999: 267–271). Precision refers to the proportion of selected items retrieved correctly by the NLP system and is reduced by wrongly selected items. Recall identifies the proportion of retrieved items (correct or incorrect) in relation to the overall amount of correct elements that should be selected by the system. These two measures are often combined into a single measure of overall performance, the F score. Statistical NLP models which typically process very large quantities of data[2] are assessed against these measures and are regarded as high quality systems with scores that may, in some cases, appear relatively low to the inexperienced observer. Studies intended to offer linguistic insight, however, may require a very high score. Here, the precision of automatic annotation and query tools is of crucial importance, and a trade-off in recall may have to be accepted, particularly in studies serving the generation of hypotheses.

Quantitative research that includes linguistic enrichment instead of working with raw data as is done in the corpus-driven research paradigm (for the latter cf. e.g. Sinclair 1991; Tognini-Bonelli 2001; and in translation studies Olohan 2004 etc.) depends on automatic annotation since the amount of text involved (the corpus used in this study counts more than one million words, see below) cannot be processed manually, particularly if the annotation is to comprise several layers. The annotation can therefore only be as accurate as the tools used (tool-related errors are systematic errors and have to be taken into account in terms of reliability). The more semantic information is included in the automatic annotation, the less accurate the tool will be. There may be applications in language technology where a comparatively low level of accuracy may be acceptable. This is, however, clearly not the case in linguistic analysis. It may therefore be advisable to employ less interpretative tools providing highly reliable results or even computer-assisted manual annotation, which may be more efficient than the manual correction of automatic annotation. Manual annotation is subject to the same limitations as interpretation in qualitative research with regard to subjectivity, inconsistency, etc. This latter aspect can be kept under control to some extent by carrying out double annotations of each text and subjecting the corpus to consistency tests (Brants et al. 2004). As to the informativity of the data, quantitative studies may have to disambiguate fuzzy sets and therefore may, under certain conditions, entail what McEnery and Wilson (2001: 77) call "a certain idealisation of the data".

As Black (1999: 6) puts it: "The pursuit of truth is desirable, but often this constitutes trying to develop a *model* of reality, an explanation of events employ-

2 Koehn's (2005) parallel Europarl corpus, for instance, contains in version 3 approx. 407m words (see http://www.statmt.org/europarl/, last visited 2 July 2010).

ing abstract and intangible concepts." This means that for the most part we cannot directly observe the things we are interested in. Consequently, we are working with hypothetical links between our abstract concepts and observable parts of reality, e.g. language, most of the time. The process of deriving observable indicators from abstract concepts is called operationalization (see chapter 4). Only these operationalized features are actually observable in texts.[3]

It is a major task of the quantitative researcher to work out the relationship between the abstract concepts and the features observed in the corpus in order to ensure the validity of the study. As mentioned in chapter 1, this relationship may sometimes be very distant, if high-level properties such as explicitness/explicitation are described on the basis of low-level features such as sentence length. In the framework of the CroCo project, this gap is reduced by adding several layers of linguistic annotation which then permit more meaningful operationalizations. Chapter 3 will discuss the linguistic enrichment in all due detail. It will also address the advantages of annotation that is not geared towards a specific theoretical framework while still allowing theory-driven queries and analyses of the corpus. This procedure will be explained in chapter 4 and exemplified in the case studies in chapters 7–11.

Finally, working with corpora of the type presented here poses an additional challenge by introducing translations as some kind of "impure" language (cf. Mauranen 2005). This is of particular relevance when translations are used to make claims about contrastive differences and commonalities in the language pair English-German. In contrastive linguistics, translations are sometimes employed as a basis of comparison to solve the problem of mapping comparable linguistic units (e.g. James 1980: 178; cf. also Johansson 2003: 35). This seems to be a somewhat adventurous approach considering the fact that translators may resort to altogether new structures not related to the respective structure in the source text when confronted with contrastive divergences. Johansson (2003: 35), however, points out that the use of balanced corpora improves the validity and reliability of this type of research (cf. also Malmkjær 1998 on corpora in contrastive linguistics and translation studies).

---

3 It has to be kept in mind, though, that categories like 'noun' and 'nominalization' are theoretical concepts again. The linguist will assign these categories to certain units in a text, but they are not "natural" features of linguistic elements. This example illustrates the theory-ladenness or the degree to which the analyst works on hypotheses like "frequent nominal elements are a symptom of an expository goal" and "linguistic units with given grammatical characteristics are nouns". All of these limit the empirical knowledge to be gained from the study of language in use, since it means that we do not simply observe and describe "brute data" (Bishop 2007: 21), i.e. data that exist without any interpretation.